DOCUMENT RESUME

ED 227 859 IR 050 154

AUTHOR Blood, Richard W.
TITLE Evaluation of Online Searches.
INSTITUTION American Library Association, Chicago, IL. Machine Assisted Reference Section.
PUB DATE Nov 81
NOTE 58p.
PUB TYPE Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Evaluation Criteria; *Evaluation Methods; Evaluation Needs; Information Retrieval; *Library Services; *Online Systems; *Program Evaluation; *Questionnaires; Records (Forms); *Reference Services; Relevance (Information Retrieval); User Satisfaction (Information)

ABSTRACT

Based on an analysis of online search evaluation forms collected from all types of U.S. libraries, and a pilot test of a draft evaluation form in selected federal research libraries, this report presents the work of the American Library Association's (ALA's) Machine-Assisted Reference Section (MARS) Committee on Measurement and Evaluation. The committee was concerned with the evaluation of searches conducted on commercially available systems such as DIALOG, ORBIT, and BRS. The report emphasizes the process of evaluating online services and the key importance of identifying clearly stated, measurable goals as the initial step in the evaluation process. Four major performance criteria--recall, relevance or precision, user effort, and response time--are identified as major evaluation criteria. The search evaluation form recommended by the MARS committee is appended. The design and use of this form and each of its 10 questions are discussed, as well as its place in an online search service's management information system. Questions ordinarily included on evaluation forms, but excluded from this recommended form, are also reviewed. Finally, the report considers sampling the users of online services, the analysis and interpretation of evaluation forms, and the implementation of modifications to online reference services based on users' evaluations. (Author/ESR)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED227859

EVALUATION OF ONLINE SEARCHES

Submitted by Richard W. Blood
San Francisco State University
to the ALA/RASD/MARS Committee
on Evaluation of Services:
November, 1981.

Reviewed and Recommended for
Publication by the MARS Committee
on Evaluation of Services and the
MARS Executive Committee at the
1982 Midwinter Conference of the
American Library Association:
January, 1982.

Note: The Appendix to this report contains
the "Search Evaluation Questionnaire"
recommended by the ALA/RASD/MARS
Committee on Evaluation of Services.

IR 050 154

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Richard W. Blood

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

EXECUTIVE SUMMARY

This report presents the work of the ALA/RASD/MARS Committee on Evaluation of Services pertaining to the evaluation of searches of computer-readable bibliographic databases--essentially online versions of printed indexing and abstracting services--available for interactive searching on commercially-available computer systems (e.g. DIALOG, ORBIT, BRS, etc.). Evaluation of online library catalogs is outside the scope of this report.

The process of evaluating online services is emphasized as is the key importance of identifying clearly stated, measurable goals as the initial step in the evaluation process. Four major performance criteria--recall, relevance or precision, user effort, and response time--are identified as the major criteria for evaluating online searches in support of the major goal for evaluation proposed by the MARS Evaluation Committee: identification of instances and causes of poor search results and user dissatisfaction as well as the modifications that might be made in any aspect of online services to increase performance levels of online searches.

Critical considerations for the design of the search evaluation form are examined. The evaluation form, moreover, is considered within the context of the entire set of forms compiled for online searches from the initial request stage through the final evaluation stage; this set of forms constitutes key components of the management information system of the online search service. The search evaluation form recommended by the

MARS Committee on Evaluation of Services is included as the Appendix to this report. Each of its ten questions is discussed. Questions ordinarily included on evaluation forms in general use in libraries, but excluded from the recommended form are also discussed as are the intended as well as the inappropriate uses of the recommended evaluation form.

This report also considers the sampling of users of online services and the need to follow up survey respondents and non-respondents. Finally, this report examines the analysis and interpretation of evaluation forms completed by users of online services and the implementation of modifications, which are suggested by analysis of evaluation data, to increase performance levels of online services.

TABLE OF CONTENTS

INTRODUCTION

FORMATION OF MARS COMMITTEE ON EVALUATION OF SERVICES

COMMITTEE METHODOLOGY

THE PROCESS OF EVALUATING ONLINE SERVICES

GOALS FOR EVALUATING ONLINE SERVICES

LEVELS OF EVALUATION

PERFORMANCE CRITERIA FOR EVALUATION

DEFINITION OF THE ONLINE SEARCH

- Definition of Reference Transaction
- MARS Measurement Committee's Definition of Online Search

INAPPROPRIATE USES OF RECOMMENDED EVALUATION FORM

- Searches of Full Text and Numeric Databases
- Searches Performed by End Users
- Known Item Searches
- Ready Reference and SDI Searches

EVALUATION FORM IN A MANAGEMENT INFORMATION SYSTEM

- Search Request Form
- Search Strategy Worksheet
- Search Log
- Search Evaluation Form

DESIGN OF THE SEARCH EVALUATION FORM

- Motivating Users to Complete the Evaluation Form
- Phrasing of Questions on the Evaluation Form

COMMENTARY ON RECOMMENDED EVALUATION FORM

- Purpose of the Survey
- Search Number
- Anonymity of Respondents
- User Identification Data
- Question No. 1
- Question No. 2
- Question No. 3
- Question No. 4
- Question No. 5
- Question No. 6
- Question No. 7
- Question No. 8
- Question No. 9
- Question No. 10

COMMENTARY ON QUESTIONS EXCLUDED FROM THE EVALUATION FORM

- User Awareness of Online Services
- Previous Use of Online Services
- Comparisons with Manual Searches
- Evaluation of the Search Intermediary
- Format of Document Surrogates

RECOMMENDATION FOR SAMPLING AND FOLLOW-UP

- Sample Size
- Sampling Technique
- Follow-up of Non-Respondents
- Follow-up of Respondents

ANALYSIS AND INTERPRETATION OF THE EVALUATION FORM

Users' Perceptions of Search Results

Interpretation of Negative Comments

Interpretation of Positive Comments

Compiling Data on Poorly-Rated Searches

Multiple Databases Searches

Comparisons of Evaluation Results

IMPLEMENTATION OF MODIFICATIONS TO ONLINE SERVICES

CONCLUDING REMARKS

Longitudinal Studies of Cost-Benefit Effectiveness

Intended Use of Recommended Evaluation Form

Survey of Non-Users of Online Services

Evaluation of Document Delivery System

ACKNOWLEDGEMENTS

REFERENCES

APPENDIX: SEARCH EVALUATION QUESTIONNAIRE

INTRODUCTION

Traditional reference services have been offered by libraries for almost a century. Key concerns relating to the provision of this service--collections, bibliographic control, training of reference specialists, etc.--have long ago been resolved. After a century of experience, however, the evaluation of traditional reference services is generally non-existent or suffers from ambiguous goals, questionable methodology, and poor survey instruments.

Similarities exist between efforts to evaluate traditional reference and online services. For less than a decade, online services have been offered by libraries, but key concerns relating to the provision of this service--hardware, database aids, training of search intermediaries, publicity, etc.--have long been resolved. As is the case with respect to traditional reference services, however, the evaluation of online services is generally non-existent or suffers from ambiguous goals, questionable methodology, and poor survey instruments.

The surge of interest in online services in the 1970's caught most reference librarians unprepared to plan for and apply online technology to the delivery of information services. Reference librarians were eager to participate, however, in the Machine-Assisted Reference Services (MARS) Discussion Group formed in 1976 within the Reference and Adult Services Division (RASD) of the American Library Association (ALA). At the Annual Conference of ALA in 1977, Peter Watson presented a

petition, with the required 250 signatures, for MARS section status within RASD to the RASD Board of Directors, which voted unanimously in favor.[1]

MARS members quickly identified measurement and evaluation as areas of key concern. As early as the third issue of Messages from MARS, the section newsletter, Sally Knapp, Chair of the MARS Section, noted growing interest in: "What statistics are being kept on computer-based reference services? What should be counted? How do you analyze what you count? How could better statistics help you to manage your services better?"[2]

Owing to declining library budgets and growing interest in accountability in the management of resources, evaluation of all library services is of current concern to librarians and the administrators of their parent organizations. Keen is the interest now shown in the evaluation of traditional reference and online services, an interest manifest in the increasing number of professional publications, ALA committees, workshops and seminars, and conference programs concerning the measurement and evaluation of library services. Within the American Library Association, interest in the evaluation of online reference services is focused within the Reference and Adult Services Division, specifically the Machine-Assisted Reference Section (MARS).

FORMATION OF MARS COMMITTEE ON EVALUATION OF SERVICES

Achievement of section status for MARS permitted formation of committees to study issues in online searching of interest to the MARS membership. At the 1978 Annual Conference of ALA an Ad Hoc Committee on Evaluation of Services was formed

to study measurement and evaluation concerns. The preliminary statement of objectives indicated that the Committee was "to develop criteria, instruments, and procedures to measure the quantity and evaluate the quality of computer-based reference services provided by libraries".[3]

After the 1978 Annual Conference of ALA, the Ad Hoc Committee was split into two separate committees: a Committee on Measurement of Services and a Committee on Evaluation of Services. The Measurement Committee was charged with studying quantitative measurements of service; the Evaluation Committee was charged with studying qualitative evaluations of service.

The specific charge made to the MARS Evaluation Committee was "to collect, analyze, and disseminate to the MARS membership information on qualitative evaluations of service and uses to which such evaluations can be put, to recommend criteria, instruments, and procedures for performing such evaluations."[4]

COMMITTEE METHODOLOGY

Five stages can be identified in the methodology used by the MARS Evaluation Committee to develop a search evaluation form. First, many evaluation forms in general use in all types of libraries were collected. Second, each sample evaluation form was analyzed as to content; questions of a similar intent such as those relating to the purpose of the search, relevancy, etc., were listed in groups to show variations in phrasing. Third, at meetings of ALA the Committee met and discussed the goals of search evaluation and debated questions that would elicit feedback from users to attain evaluation goals. A preliminary version of the search evaluation form was drafted by the Committee for use in a pilot test of its effectiveness. Fourth,

a pilot test of the draft evaluation form was done in selected federal research libraries. From managers of online services participating in the pilot test, comments were received regarding their experience with use of the evaluation form. The pilot test generated suggestions for improvement, especially in the sequence and phrasing of questions and response categories. Fifth, based on the feedback from pilot test participants, the evaluation form was revised. Input on the revised evaluation form was further solicited at the 1980 Annual Conference of ALA during a joint hearing of the MARS measurement and evaluation committees, where comments and recommendations for improvement were received from a cross section of interested librarians. This feedback was analyzed and used to draft the final version of the search evaluation form, which appears as the Appendix to this report.

THE PROCESS OF EVALUATING ONLINE SERVICES

The recommended search evaluation form that appears as the Appendix to this report must be understood within the context of the total process for evaluating online services. This process consists of five sequential steps. First, the goals of evaluation must be identified, clearly stated, and measurable. The "why to evaluate" and the "what to evaluate" suggest operational aspects of the "how to evaluate;" that is, the goals of evaluation guide the design of the evaluation form and the procedures for the evaluation. Evaluation forms in general use in libraries suffer from ambiguity in evaluation goals.

Second, after the goals for evaluation are clearly stated, the evaluation form and methodology must be designed to collect, within cost constraints of the program, the evaluative feedback

from users needed to assess progress toward the goals of the evaluation program.

Third, the actual evaluation program must be conducted. Evaluation must not be regarded as a one-time study, but rather as a continuing program for monitoring quality in the delivery of online services. Evaluation must be undertaken on a continuing basis to measure the effects of changes in service policies and procedures.

Fourth, evaluation results must be analyzed and interpreted to ascertain causes of poor search results and modifications that might be made in any aspect of the service to expedite attainment of evaluation goals.

Fifth, the modifications suggested by the analysis and interpretation of evaluation results must be implemented. The process of evaluating online services consists of not fewer than these five steps. The evaluation program is merely an intellectual exercise if it fails to implement modifications to improve the overall performance of online services.

GOALS FOR EVALUATING ONLINE SERVICES

Controversy and confusion in the literature on the evaluation of online services stem from the lack of explicity defined goals. Ladendorf observes:

> ...without careful goal definition, there is no way to perform intelligent evaluation. Also, measuring progress toward different kinds of goals usually demands the use of different kinds of evaluative measures. Much of the controvery in the literature on evaluation has been created by an underlying disagreement about goals. Endless debates occur on what should be

measured and how it should be measured. Such arguments are futile unless the protagonists are interested in achieving the same goals. These goals must be defined before good evaluative measures can be worked out.[5]

Evaluation of online services should produce two types of data: "performance figures for a representative group of searches" and "examples of system failures to allow analysis of causes of failure."[6] F.W. Lancaster argues that evaluation:

> involves analytical procedures whereby the major sources of system failure are identified, thus allowing corrective action to be taken to raise the performance level of the system. . . .
>
> Evaluation is essentially a diagnostic procedure which, like other forms of diagnosis is intended to lead to therapeutic action. An evaluation program, hopefully, is not conducted merely as an intellectual exercise. Thorough evaluation tends to be expensive, and can only be justified if the evaluation program is likely to lead to significant improvement in the performance of the system.[7]

The Evaluation Committee feels that the primary goal of evaluating online services is to identify and analyze instances of poor search results and user dissatisfaction and to ascertain the modifications that can be made to improve the performance of online services. The Committee also recognizes, however, that managers of online services might have other administrative goals for surveying users of online services; one of these administrative goals, for instance, might be the need to obtain data from

users to justify budgetary and staffing resources committed to the online program. In the interest of keeping the recommended evaluation form as short and simple as possible to encourage good response rates from users, the questions included reflect only diagnostic goals for evaluation. Managers of online services should include whatever other questions on their evaluation forms that are needed to attain other administrative goals for evaluating online services. Chapter three of Don Dillman's Mail and Telephone Surveys (New York: Wiley, 1978) discusses general principles for writing questions for survey questionnaires.

LEVELS OF EVALUATION

F.W. Lancaster defines three levels of evaluation: effectiveness, cost-effectiveness, and cost-benefits evaluation. Effectiveness is defined as "evaluation of system performance in terms of the degree to which it meets user requirements." Cost-effectiveness is defined as "evaluation in terms of how to satisfy user requirements in the most efficient and economical fashion." Cost-benefits evaluation is defined as "evaluation of the worth of the system(i.e., is the system worthwhile, does it justify its existence?)."[8] Owing to methodological problems in measuring costs in different library environments and benefits to users from online services, the Committee's work emphasizes evaluation at the level of effectiveness.

PERFORMANCE CRITERIA FOR EVALUATION

Four major dimensions of the effectiveness of online searches are commonly identified: recall, precision or relevance, user effort, and response time.[9] These dimensions of evaluation

constitute performance criteria that should be measured by users of online services. Most of the questions included on the recommended search evaluation form elicit feedback of diagnostic value regarding these four performance criteria.

DEFINITION OF THE ONLINE SEARCH

Online searching may be defined as "the process of interactively searching for and retrieving information by computer from a machine-readable file (database) of information."[10] In the literature of online searching, confusion abounds on the definition of the "online search," with two contradictory views prevalent. One view regards the online search as "one interactive computer access to one database"; the other view regards the online search "as many interactive accesses by computer to as many databases as the searcher considers necessary to conclude the search."[11] In deciding which definition to recommend for use with the search evaluation form, the MARS Evaluation Committee considered both the definition of a reference transaction recommended by the ALA/LAMA Committee on Statistics for Reference Services and the definition of the online search recommended by the MARS Measurement Committee.

Definition of Reference Transaction

An online search is a reference transaction; the definition of the online search, therefore, should be consistent with--or at lease not incompatible with--the definition of a reference transaction. The Committee on Statistics for Reference Services recommends that a reference transaction be defined as:

> an information contact which involves the use, recommendation, interpretation, or instruction in the use of one or more information sources, or knowledge of such sources, by a member of the reference/information staff.[12]

In accord with this definition, accesses to multiple information sources (e.g. consultation of three printed abstracting services) required in the search to satisfy one intellectual query are not counted as separate transactions, but as part of the same reference transaction.

MARS Measurement Committee's Definition of Online Search

The MARS Measurement Committee defines the online search to make it compatible with the preceding definition of a reference transaction. It defines the online search as "the interactive access by computer to as many data bases as the searcher considers necessary to conclude the search.[13] In accord with this definition, "an online search accessing five databases to answer one query would be counted as one search, not five.[14] Since the MARS Evaluation Committee and the MARS Measurement Committee agreed to cooperate closely, not to duplicate each other's work, and to follow, wherever possible, each other's recommendations, the MARS Evaluation Committee accepts this definition of the online search for use with the search evaluation form.

If access to three databases are required to process a single search request, for example, the user would ordinarily receive a search output consisting of document surrogates (e.g. citations, abstracts) from three databases, but should be given only one evaluation form and be asked to evaluate overall the combined search output from the three databases, instead of

being asked to evaluate separately the search results from each of the three databases.

INAPPROPRIATE USES OF RECOMMENDED EVALUATION FORM

The recommended evaluation form, however, cannot be used to evaluate all types of online searches consistent with the definition of the online search as recommended by the MARS Measurement Committee. The recommended evaluation form should be used primarily to evaluate retrospective searches of document surrogate databases (e.g. ERIC, BIOSIS, etc.) performed by a search intermediary to satisfy a request for information involving unknown items.

Searches of Full Text and Numeric Databases

With questions framed primarily to evaluate document surrogates, the recommended evaluation form is not intended for use in evaluating searches of full text databases. Similarly, the evaluation form is not intended for use in evaluating searches of numeric or statistical databases, where often the objective is either the retrieval of specific numeric data or sets of numeric data for use in various statistical manipulations (e.g. factor analysis, regression and correlation analysis, etc.) also performed online. A form appropriate for evaluating searches of numeric databases would not contain questions relating to recall and precision in the traditional sense, but would instead emphasize capabilities for retrieving specific numeric data and mathematically manipulating the data as well as the user cordiality of these capabilities.

Searches Performed by End Users

Neither is the recommended evaluation form intended for use in evaluating searches performed by end users; the form should not be used, therefore, to evaluate direct use by end users of conventional search systems (DIALOG, SDC, BRS, MEDLINE) or of bibliographic utilities such as OCLC, RLIN, WLN, or UTLAS. Neither is the form intended for use in evaluating direct use by end users of local or regional online catalogs.

Known Item Searches

Since a search executed to retrieve a known item (e.g. a specific technical report or patent, the publications of a personal or corporate author, etc.) either retrieves or fails to retrieve the desired items(s), traditional concepts of recall and precision fail to apply as in searches to recrieve unknown items. In some cases, a known item search strategy is used to search for unknown items: for example, searching with key words from the name of an organization known to be involved in research on an esoteric subject for which topical terminology is still in flux. If the intent of the search is to retrieve unknown items, the recommended evaluation form can still be used to evaluate the search regardless of the search strategy used by the search intermediary.

Ready Reference and SDI Searches

The recommended evaluation form is appropriate neither for evaluating ready reference nor selective dissemination of information (SDI) searches. The recommended form emphasizes traditional performance criteria of recall and relevance, an

inappropriate emphasis for searches done for ready reference queries in which the objective is the retrieval of a few key citations or a citation to a work containing the answer to a specific query. Since SDI services involve searches performed at intervals to satisfy a current awareness information need, a form of significant diagnostic value for evaluating SDI searches must include questions on terms, phrases, corporate authors, etc. producing many "false drops" or largely irrelevant citations to improve precision levels of future SDI searches.

EVALUATION FORM IN A MANAGEMENT INFORMATION SYSTEM

Managers of online search services require systematic collection of information, using an integrated system of forms. The major forms used in the management of online search services are: search request form, search strategy worksheet, search log, and search evaluation form. Each form is designed to compile data for a particular use at a successive step in the progression of a search from the presentation of the request to the provision of evaluative data on the results of the search by the user.

Search Request Form

The search request is typically the first form generated in the systematic collection of information on a search. The search request is completed at the time the user submits a search request or at the time the search intermediary interviews the user to negotiate the search request. This form is completed by the user, or jointly by the user and the intermediary working in consultation, or by the intermediary in

response to verbal queries directed in person or over the telephone to the user.

The major objective of the search request is to record enough pertinent information about the search request to ensure a productive search. The search request usually includes: (1) a statement of the topic and purpose of the search in the words of the user, (2) a list of key terms, phrases, codes, personal and corporate authors, etc. to be included or excluded, (3) qualifiers relating to type of document surrogates (e.g. citations, abstracts, etc.) desired with respect to dates of coverage, languages, format (e.g. technical report, patent, conference proceeding, journal article, etc.), and (4) a statement of the user's preference as to the type of search desired, especially with respect to the tradeoff between recall and relevance at various levels of comprehensiveness. In addition to collecting data needed to execute a productive search, often the search request form collects data relating to general aspects of the online service such as how the user found out about the service.

Search Strategy Worksheet

The search strategy worksheet is the next form generated in the processing of a search request. Some online services leave space on the search request form to formulate search strategy, obviating the need for a separate form. Sometimes a search strategy worksheet is used only to formulate search strategy for complex searches, with scratch paper or no paper

whatsoever used to formulate relatively simple search strategies.

The major objective of the search strategy worksheet, or whatever is used in lieu of it, is to assist in the formulation of effective search strategy. The search strategy worksheet includes lists of approved terms, phrases, codes, and names checked in various thesauri and database aids, lists of key qualifiers, and the set of search statements to be submitted to the system as the initial search strategy complete with terms, phrases, codes, and qualifiers appropriately linked with Boolean operators. Sometimes alternate search strategies are formulated in advance and recorded on the worksheet to allow the search strategy to be easily broadened or narrowed in response to intermediate search results.

Search Log

The search log is usually the next form completed in the processing of search requests. At the end of the online session, the search log is completed by the search intermediary, who makes a separate entry for each database accessed on each search system. If a search request required five databases to be accessed, for instance, five entries--one entry for each database accessed--are made in the search log.

The major objective of the search log is to collect cost and other data expediting the reconciliation of invoices received from search system vendors. The search log may include the numbers or names of the databases accessed, indication of the search system used, connect time, number of offline or online prints, and costs. An example of a search log appears

in a recent issue of the journal Online.[15]

Search Evaluation Form

The search evaluation form is generally the last form related to the processing of a search request. Since the goals of the search request, the search strategy worksheet, and the search log are clear and fairly similar from one library to another, the content of these forms is similar from library to library. In marked contrast to the goals for these three forms, the goals for evaluating searches are often ambiguous and not explicitly stated; the content of the evaluation form, therefore, predictably varies more from one library to another than in the case of the other forms. In relation to the great ambiguity and confusion in the goals for evaluating online searches, however, the search evaluation forms in common use are more similar than would be expected--probably the result of libraries copying or modeling their evaluation forms on those of other libraries.

DESIGN OF THE SEARCH EVALUATION FORM

Unlike the other three major forms, the search evaluation form is rarely completed by or with the assistance of the search intermediary, but instead by the user following examination of search results consisting of document surrogates. Two major considerations for the design of the search evaluation form stem from its completion by the user, rather than by the search intermediary.

Motivating Users to Complete the Evaluation Form

First, users must be sufficiently motivated to take time to read, complete, and return the evaluation form. Motivating

users to complete or assist in the completion of the search request form is not ordinarily a problem, because this form is completed at a time when users still have an information need that is expected to be satisfied by the use of the online service. In other words, users regard completion of the request form as simply one of the necessary efforts to use the service. Motivating users to complete the evaluation form, however, is a major problem, because the form is completed after the information need has been presumably satisfied; consequently, satisfaction of the information need is not affected by failure to complete the evaluation form.

Users can be motivated to read, complete, and return the evaluation form by keeping it as short and simple as possible. A long evaluation form intimidates users and reduces response rates and increases the effort of users in using the service; this consideration is germane, for user effort is consistently identified as a major performance criterion in the evaluation of information services.

Since the evaluation form is only one form of an integrated set of forms used to collect different data at successive stages in a search, all the data required to analyze and interpret responses on the evaluation form and ascertain causes of failure in the service need not be collected on the evaluation form. A search number should be assigned to each search request form and used on all subsequent forms--search strategy worksheet, search log, and evaluation form--to link together the total data compiled on a search. If the evaluation form does not provide diagnostic data sufficient to analyze the cause of a poorly-rated search, the search linkage number recorded on the evaluation form

can be used to locate other search forms completed for the same search to allow examination of other data that may help ascertain the cause of poor performance and system failure. For this reason, most of the questions recommended for the search evaluation form focus on the user's evaluation of search results, especially in relation to the major performance criteria.

Users can also be motivated to complete and return the evaluation form by making its completion and return as convenient as possible. Fixed response categories are recommended for most questions, allowing users simply to check the category that best describes their response. The form should allow for easy return to the search service by including a pre-addressed envelope or by making the form self-mailing. If return of the evaluation form through the U.S. mail is required, the envelope or self-mailing form should be pre-franked.

Phrasing of Questions on the Evaluation Form

The second design consideration stemming from completion of the evaluation form by the user, rather than by the search intermediary, is the need to phrase questions appropriately, with special emphasis on using language understandable to most users. The search evaluation form must be free from the jargon of information science and online searching; terms such as "document surrogates", "recall ratio," and "false drops," are meaningless to most users and must not be used on the evaluation form.

COMMENTARY ON RECOMMENDED EVALUATION FORM

The search evaluation form recommended by the MARS Evaluation Committee appears as the Appendix to this report.

Developing an evaluation form and presenting it as a recommendation, without accompanying commentary, would fall short of the Committee's charge. This approach would also deprive others of the valuable discussions regarding the selection and phrasing of questions on the evaluation form. The following commentary does not record the full debate on each question, but it is sufficiently complete to reflect the consensus rationale used to select and phrase each question.

Purpose of the Survey

Users are more likely to complete and return a questionnaire when informed of the purpose of the survey; at the head of the evaluation form appears, therefore, the statement: "In order to evaluate and improve computerized literature search services, we would appreciate your taking a few moments to complete and return this questionnaire." The name or acronym of the online service may be substituted for "computerized literature search service". To this basic statement eliciting user cooperation, other statements may be added to encourage users to respond to the questionnaire or to provide local instructions on the return of the questionnaire.

Search Number

Staff of the online service must record the "search number" on the evaluation form before it is distributed to the user with search results. The search number must match the number assigned to the search on the corresponding request form and subsequently recorded on all additional forms completed for the search. The search number uniquely identifies each online search and links all search forms completed for a search,

making available for analysis and interpretation the total data compiled on a search from the request to the evaluation stage. The search number is also used to eliminate anonymous respondents from the telephone follow-up of non-respondents.

Anonymity of Respondents

If some users want to remain anonymous, they may be given the option of withholding their names on the evaluation form. This option will, in some cases, remove any threatening aspect the respondent may perceive. Inclusion of a unique number on a questionnaire does not permit the respondent to remain completely anonymous, for the search number can be used to reveal identity by obtaining the respondent's name from the search request form bearing the same search number. Since each evaluation form must contain the search number, a statement needs to be added indicating that the search number is used to identify non-respondents for subsequent follow-up by telephone and is not used to reveal the identity of respondents returning questionnaires completed anonymously. To encourage users to be completely frank in completing the evaluation form, a statement is also added that promises to keep responses "strictly confidential" whether or not respondents choose to remain anonymous.

User Identification Data

The next section on the evaluation form asks for user identification data: name, address, telephone number, and status. The telephone number is particularly important to permit telephone follow-up of users whose responses on the evaluation form are apparently contradictory, ambiguous, or otherwise difficult to analyze and interpret. Although the

telephone number can be obtained from the corresponding request form, inclusion of the telephone number on the evaluation form permits telephone follow-up without referring to the corresponding request form. Users are also asked to check a response category indicating their status. Since the status categories of users vary by type of library, an online service should list categories that cover most potential users. The status of the user is included on the evaluation form, despite its frequent appearance on the request form, because of its diagnostic value in interpreting evaluation results. Some user identification data recorded on the search request form are repeated on the evaluation form to expedite analysis and interpretation of evaluation results by limiting the need to refer to the original request form. The rest of the commentary discusses each of the ten questions included on the recommended evaluation form; question numbers correspond to those on the recommended evaluation form included in the Appendix to this report.

Question No. 1

> What was your main purpose in requesting this search? In other words, at the time you submitted your search request, what had you planned to do with the results?

The evaluation form emphasizes evaluation at the level of effectiveness--that is, evaluation of search results by the user in the context of the purpose for which the search was requested. Consequently, a response to this question is essential to interpret responses to other questions.

A question regarding the purpose of the search is often included on the search request form to help formulate effective

search strategy; this question is also appropriate for the evaluation form to expedite analysis of responses to the rest of the questions without having, in every instance, to refer to the original search request form through use of the search linkage number.

Another reason for including this question on the evaluation form is that it requires users to think about the purpose of the search before responding to any other questions. By asking this question first, the appropriate context is established for evaluation of search results--that is, this question encourages users to evaluate search results and answer the rest of the questions in the context of the purpose for which the search request was made.

Listing response categories appropriate for all types of libraries and users would needlessly extend the list of response categories. Response categories listed for this question should be locally phrased and include the purposes responsible for most of the search requests submitted to a specific online service. Response categories for the academic library might include: "research for an undergraduate term paper," "research for a graduate seminar paper," "research for an M.A. or M.S. thesis," "research for a doctoral dissertation," "research for a grant proposal," "individual faculty research," etc. Response categories for the public library would differ from those appropriate for an academic or special library. The response categories can be made as inclusive as possible, without forcing the user to respond by checking a related, but not entirely appropriate, category by including the category "Other? (specify)_____" as the last response option for

this question.

Question No. 2

> Was the purpose of this search to determine that no previous work had been done on this topic?

Most users submitting search requests want to retrieve document surrogates relevant to a purpose. Some users, however, submit search requests to determine that no previous work has been done on the topic of the request. Users investigating topics for doctoral dissertations, users investigating topics on the "research frontier" or at the "cutting edge of scientific knowledge," and users searching the existence of a patent often request searches in the expectation--and in the hope--of retrieving no published work. This question is included to avoid misinterpretation of apparently "negative" search results noted by the respondents.

Question No. 3

> Does this search provide enough relevant citations for the purpose for which you submitted the search request?

This question asks whether or not the level of recall of the search results is adequate for the purpose for which the search request was made. Although a few users need a comprehensive search retrieving as many relevant citations as possible, most users need search results that provide enough relevant citations for the purpose for which the search request was made. Most users have an intuitive feeling for the number of relevant citations that they require to accomplish their purposes within time and cost constraints.

The response category "No, but didn't expect to see anything" is included to expedite analysis of data from users submitting search requests to determine that no previous published work had been done. The "please comment" annotation is attached to the simple "No" response to encourage users to comment when not enough relevant citations are retrieved. It is anticipated that some users responding "No" to this question can provide commentary of diagnostic value that can be used to improve recall levels of future searches.

Question No. 4

> Among the total citations provided by this search, what percentage appears relevant to the specific question or topic for which you submitted a search request?

A measure of recall by itself is meaningless without an accompanying measure of relevance or precision, for total recall can always be achieved by simply printing out the contents of the entire database. Questions regarding relevance or precision measure the filtering power of the search strategy--that is, the capability of the search strategy to retrieve enough relevant citations and simultaneously prevent retrieval of too many "false drops" or citations not relevant to the purpose of the search. Although "precision" terminology has replaced "relevance" terminology in discussions of relevance in information science, the term "relevant" is used in this and other questions, because to users "relevant" is a more meaningful term than "precision."

This question asks users to judge the relevance of the search results with respect "to the specific question or topic" for which the search request was made and not with respect to

the overall information need. The output of any online search can be fairly evaluated only in terms of the specific question posed for retrieval.

Five response categories are included for this question. The first category "0%" is included to accommodate a specific response when search results do not provide any relevant citations. The remaining four response categories are ranges of uniform size, are mutually exclusive, and account for the entire range of possibilities. Inclusion of more than five response categories would increase users's efforts in responding to this question without adding much useful diagnostic data. Also, users may not be able to estimate the sufficiency of search results more precisely.

Question No. 5

> Among the total citations provided by this search, what precentage appears relevant to your overall information need, rather than simply relevant to the specific question submitted as a search topic or question?

This question asks users to judge the relevance of the search results not just with respect to the specific topic of the search request, but in terms of the overall information need; it asks users to broaden judgment of relevance. These two concepts of relevance--relevance to the specific search topic and to the overall information need--are quite distinct. Citations not relevant to the question posed in the search request may be relevant to the overall information need. Some users will probably fail to understand the distinction between this and the preceding question, a misunderstanding, which might, unfortunately, require telephone follow-up to clarify

the distinction sufficiently to permit valid analysis.

As noted in the commentary on the previous question, relevance to the specific question posed for retrieval must be the basis for judging the performance of online searches, not relevance to the overall information need. Citations not directly relevant to the specific question posed for retrieval, but still relevant to the overall information need, nonetheless, contribute to the satisfaction of users; the response to this question may be useful, therefore, in analyzing and interpreting responses to other questions on the evaluation form.

The five identical response categories included for the previous question are included for this question to ensure comparability between responses to these two questions measuring two valid concepts of relevance. If users understand the distinction between this and the previous question, the response category checked for this question should be in a percentage range as high or higher than the percentage range checked as the response to the question concerning relevance to the specific topic of the search request, because relevance to the overall information need is a more inclusive concept than is relevance to the specific topic of the search request. In other words, citations relevant to the search topic are also relevant to the overall information need, but citations relevant to the overall information need may not be relevant to the specific question posed by the search request.

Question No. 6

> Among the relevant citations provided by this search, what percentage is new to you, or, in other words, was unknown to you at the time you examined the search results?

This question asks users to judge the novelty of the relevant citations retrieved by the search by asking for the percentage of the relevant citations that were previously unknown. An online search cannot be fairly evaluated on the basis of whether or not the relevant citations retrieved were previously unknown, but the retrieval of relevant and previously unknown citations is of primary importance to users and is expected, therefore, to affect significantly the overall evaluation of search results. The five response categories included for this question are comparable to those included for the two previous questions, and the same comments regarding them pertain here in the context of this third measure of relevance that incorporates the concept of novelty.

Question No. 7

> Do you feel that the citations that are both relevant and previously unknown to you are worth the cost that you paid for the search?

This question elicits a general "Yes" or "No" response as to whether or not the relevant and previously unknown citations are worth the cost paid for the search. F.W. Lancaster refers to this aspect of an online search as the "novelty-cost ratio."[16] Obtaining an adequate number of relevant and previously unknown citations at a reasonable cost is the goal of most users in submitting requests for online searches. Users that check "No" as the response to

this question are further asked to "please comment" in the hope of eliciting comments of diagnostic value. If users are not charged for online searches, this question should be omitted from the evaluation form, or modified to read: "worth the time and effort you spent on the search."

Quesition No. 8

> Was the time lapse between submitting your search request and receiving your search results reasonable?

This question elicits an overall "Yes" or "No" judgement as to whether or not the time lapse was reasonable between submission of the search request and the receipt of the search results. This question is germane to the evaluation of online services, because the response time of information retrieval systems is consistently identified in the literature as a major performance criterion in evaluation. Users that check "No" as the response to this question are further asked to "please comment" in the hope of eliciting comments of diagnostic value regarding timeliness in the delivery of online services.

Question No. 9

> Were the results of the search of value to you?

This broad question elicits a comprehensive judgment on the overall value of the search results. The response to this question is often of key value in analyzing and interpreting responses to more specific questions. Responses to more specific questions may seem to indicate a positive or negative evaluation in isolation, but that interpretation must be related to the overall evaluation of search results. Users that

check "No" as the response to this question are further asked to "please comment," again in the hope of eliciting comments of diagnostic value.

Question No. 10

> The major reason for the Search Evaluation Questionnaire is to obtain your comments and suggestions for improving the computerized literature searching service. If you have suggestions as to how any aspect of the search services can be improved, please comment in detail.

While the nine preceding questions elicit comments regarding search results, this broad question elicits comments and suggestions regarding any aspect of the online service. Deliberately broad and non-suggestive, this question encourages users to comment on any aspect of the service needing improvement. In response to this question, users can comment on aspects as diverse as the pricing policy for online services, the value of the user's presence or absence during the online session, the desirability of offline v. online printing of search results, the assistance received from the search intermediary, etc. This question elicits comments from users that may be of diagnostic value to the criteria for evaluating online services.

COMMENTARY ON QUESTIONS EXCLUDED FROM THE EVALUATION FORM

The standard for inclusion of a question on the recommended evaluation form is whether or not it is likely to elicit meaningful and useful responses indicative of the user's evaluation of search results in terms of four major performance criteria--recall, relevance, user effort, and response time--that would suggest user satisfaction or dissatisfaction with online services and, most importantly, identify instances of

of search failures. Poorly-rated searches are then to be analyzed to determine whether or not modifications are needed in policies, procedures, hardware, search strategies, training of search intermediaries, etc. to increase overall performance levels of future searches. Another requirement for a question to be included on this evaluation form is whether or not it can better be asked on another search form, especially the search request form. The intent is to keep the evaluation form as short as possible to minimize the effort of users in complying with evaluation and to encourage a good return rate for the questionnaire.

Each type of question found on evaluation forms in general use in libraries was discussed as a potential candidate for inclusion on the recommended evaluation form. The following commentary explains why some questions found on evaluation forms in general use are excluded from the recommended evaluation form.

User Awareness of Online Services

A question excluded from the recommended evaluation form asks how the user found out about online services. This question is meaningful, for it provides data to evaluate alternate channels for publicizing the availability of online services. This question needs to be asked on the search request form, however, because users are more likely to remember the communications medium responsible for finding out about online services at the time the search request is made, rather than at a later time when the search results are evaluated.

Previous Use of Online Services

Another question excluded from the recommended evaluation form asks users about previous use of online services. Since the expectations that users have of any information system are conditioned by their past experience with the system, experienced users of online services are likely to set different standards of performance than new users. This question is meaningful, therefore, in interpreting evaluative feedback from users, but it can easily be asked on the search request form.

Comparisons with Manual Searches

Another question excluded from the recommended evaluation form asks users to compare one or more aspects of the online search to a hypothetical manual search done on the same topic. Most commonly, users are asked about how the number of relevant citations produced by the online search would compare with the number produced by a manual search or about how much time was saved by having an online search made in lieu of having done the search manually. These questions lack diagnostic value relative to the criteria for evaluating online services. Another reason for excluding these questions is that, in most instances, an online search obviates the need for a manual search on the topic, making these questions meaningless, for users lack any basis on which to answer them.

Evaluation of the Search Intermediary

Another question commonly asked on evaluation forms in general use, but excluded from the recommended evaluation form asks users to evaluate one or more aspects of their interaction with the search intermediary; most commonly, this type of question asks users to evaluate the assistance received from

the search intermediary. Many attributes of the search intermediary significantly affect the quality of search results and the level of satisfaction with online services; however, most users lack expertise to evaluate assistance received from the search intermediary. Search intermediaries are best evaluated by supervisors and peers as well as through use of self-evaluation techniques. The last question included on the recommended evaluation form is broad enough, however, to allow users to comment on the assistance received from the search intermediary.

Format of Document Surrogates

Another type of question excluded from the recommended evaluation form asks about some aspect of the format of the document surrogates retrieved by the online search; most commonly, this question asks users to evaluate the assistance provided by citations or abstracts in judging the relevance of published works to the purpose of the search. Responses to this question are of diagnostic value in instances of one database access to satisfy a search request, where users receive only citations or only citations with abstracts. With the increasing number of multiple database accesses made to satisfy a single search request, however, responses to questions regarding format of document surrogates are difficult to interpret owing to the different levels of information provided by document surrogates retrieved from accesses to multiple databases.

RECOMMENDATION FOR SAMPLING AND FOLLOW-UP

Most online services that collect evaluation data on search results survey users on a census basis, with an evaluation form

included with results for each search. Typically, only between twenty and forty percent of these evaluation forms are returned. Thus, not only is the response rate low, but also the respondents are a poor cross section of online users. Without intensive follow-up of non-respondents, surveys are often biased by the tendency for users having extreme reactions--extremely favorable or unfavorable--to be more likely to respond to questionnaires than other users. If dissatisfied users are more likely to complete and return evaluation forms than satisfied users, this bias in response rates is desirable, for the major goal of evaluation is to identify instances of poor search results to determine whether or not any aspect of the service needs modification to improve overall performance. Respondents are rarely contacted, moreover, in a follow-up to determine whether or not their responses are biased by non-sampling errors stemming from misinterpretations of questions and other factors.

Sample Size

In order to determine an adequate sample size for surveying users of online reference services, managers of online services must specify desirable limits of sampling error that can be tolerated. Standard equations can then be used to estimate the sample size needed to achieve the desired level of sampling precision. These equations require, however, the estimation of certain unknown properties of the total population of online searches. William G.Cochran's Sampling Techniques (New York: Wiley, 1963) discusses the estimation of sample size and provides the standard equations.

Sampling Technique

Online searches can be selected for inclusion in the sample based on numbers selected from a table of random digits and matched against numbers assigned to search requests. Selecting a systematic sample--every n th search for inclusion in the sample--is easier than selecting a random sample. If the online service processes about 1200 searches annually and needs a total annual sample of about 300 searches for the evaluation survey, then an evaluation form should be included with the results of every fourth search. Except in the unlikely case of repeatability in the type of user (e.g. every fourth user is always a graduate student), systematic sampling is an adequate method of sample selection.

Follow-up of Non-Respondents

If sampling is used, telephone follow-up of non-respondents and some respondents is highly recommended. Non-respondents need to be contacted by telephone to collect the data or to encourage them to complete and return the search evaluation form. If too much time elapses between examination of search results and completion of the evaluation form, the user's imperfect recollection might seriously bias evaluation responses; therefore, users must be contacted by telephone follow-up for failure to respond soon after receipt of search results.

Follow-up of Respondents

Some respondents also need to be contacted by telephone to ensure quality control with respect to non-sampling errors; this follow-up should minimize introduction of significant errors in responses owing to undetected, systematic misinterpretations of questions and response categories. Some respondents need to

be contacted by telephone, moreover, as part of the diagnostic analysis, especially in cases of poorly-rated search results in which neither data on the evaluation form nor on related search forms suggest the cause of poor search results and user dissatisfaction.

ANALYSIS AND INTERPRETATION OF THE EVALUATION FORM

Diagnosis of problems and failures in the delivery of online services requires a careful analysis and interpretation of evaluation responses. No single response should be interpreted in isolation apart from other responses. In a questionnaire survey of MEDLINE users, Tagliacozzo finds "contradictions between an overall appraisal of the service and more specific responses on the outcome of the search."[17] This finding, according to Tagliacozzo, stresses

> the importance of not trusting an overall single response by the user to represent his degree of satisfaction. Before inferring that the information needs of the users were satisfied by the service, one should access and compare their responses to various aspects of the search outcome.[18]

Users' Perceptions of Search Results

Evaluation results make sense only if one realizes that responses on search evaluation forms reflect users' perception of search results, which often correlates imperfectly with the technical performance of the information retrieval service with respect to recall and precision measures. From experience

surveying MEDLINE users, Tagliacozzo concludes:

> When a questionnaire has been used . . . one is puzzled by the contradictory and dissonant responses given to its various items. The picture which emerges from the data that we analyzed suggests that a great variety of factors, some related and some unrelated to the performance of the information retrieval system, influence the users' responses. Factors related to the measuring instrument, such as the different interpretation of questions, the tendency to avoid extreme positions on rating scales, and the questionable reliability of memory in reporting past events certainly affect the users' responses. More directly related to the service provided by the information retrieval system are the users' attitudes and perceptions, their knowledge and expertise in the area of the search, their previous acquaintance with the literature. Finally, the type of search requested by the users may be responsible for some of the variance in their judgment: certain searches by their very nature have a higher probability than other searches of producing results that satisfy the users (for example, generic searches versus specific searches).[19]

Interpretation of Negative Comments

Despite the retrieval of many relevant citations, negative judgments are sometimes made on the overall value of searches for reasons unrelated to the technical performance of search

results in terms of recall and precision measures.

Sometimes negative judgments are unfairly made as to the overall value of searches, because the relevant citations were previously known and, therefore, judged to be of little or no value. An information retrieval system must respond to the needs of all users--the novice with no previous subject knowledge as well as to the subject expert with extensive bibliographic knowledge; consequently, information retrieval systems cannot be fairly evaluated on the basis of whether or not the relevant citations retrieved were previously known to users. Negative judgments sometimes stem, moreover, from users' overly high expectations regarding online technology.

Despite good recall and precision measures, negative judgments are sometimes made on the overall value of searches, because relevant citations were missed; for example, searches may be unfairly evaluated for failure to retrieve key conference proceedings when the database accessed excludes this publication format.

Negative judgment may also be made on the overall value of searches, because original works referenced by search results cannot be located. This situation represents a failure of the local collection, interlibrary loan services, and other aspects of the document delivery system--not a failure of the online service. Users may not distinguish, however, between the performance of the online retrieval service and the document delivery system.

Interpretation of Positive Comments

On the other hand, despite poor recall and precision measures, positive judgments may also be made on the overall value of searches. Retrieval of only a few key relevant citations might elicit positive overall evaluations of search results. Sometimes "even a single reference--especially if it is the crucial one--can make a user happy."[20] Mention has been made of the possibility that users--especially those who request searches to determine that no previous work had been done on a topic--are satisfied by the failure to retrieve any relevant citations. In this situation, negative responses to specific questions on recall and relevance must not be interpreted as expressions of judgment on the overall value of the search. Tagliacozzo feels, moreover, that:

> the judgment of some users reflects their general opinion of the potential usefulness of the system rather than the specific appraisal of the service rendered to them. Many users do not want to discourage a new service, even if it has not helped them, in view of possible future advantages they might derive from it . . . . It may happen that the user's response is an expression of his general attitude toward technological innovations, rather than an appraisal of realistic and relevant factors, such as the extent to which the outcome of the search has met his needs for information.[21]

Compiling Data on Poorly-Rated Searches

In some instances, the entire set of responses on the search evaluation form cannot be properly interpreted without consulting other forms compiled for the search. Sometimes the cause for a poorly-rated search cannot be ascertained by examining the entire set of search forms; this situation requires telephone follow-up of respondents. Managers of online services must concentrate on exceptional cases; consequently, every effort must be made to marshall as much evidence as possible for poorly-rated searches from the forms compiled as searches progress from the request stage to the evaluation stage, with use of post-evaluation follow-up by telephone of respondents in which the evidence is ambiguous, contradictory, or fails to suggest the cause of poor search results or user dissatisfaction. Sampling evaluations of search results is futile unless the effort yields information of diagnostic value to ascertain causes of failure in the delivery of online services.

Multiple Database Searches

Owing to the increasing number of computerized databases, searching multiple databases to satisfy a single search request is becoming increasingly common. Hawkins and Brown, for instance, report that the number of databases accessed per search increased from an average of 1.3 in 1974 to 2.4 in 1978 at Bell Laboratories and from an average of 1.3 in 1974 to 3.6 in 1978 at the National Oceanic and Atmospheric Administration.[22]

The definition of the online search as computer access to one or more databases to satisfy one intellectual query accepted by the Committee for use with the recommended evaluation form might cause problems in analyzing and interpreting responses to

some questions. These problems stem from asking users to evaluate overall citations retrieved from multiple databases. It might be difficult to interpret responses to the questions asking for the percentage of relevant citations, for this percentage might vary considerably from one database to another. The response to this question requires a judgment on the overall percentage of relevant citations retrieved from all the databases accessed for the search. Information on the percentage of relevant citations for each database accessed is lost, therefore, to the search analyst.

Since the number of users asked to complete evaluation forms can be limited by sampling, the time saved by including fewer users in the evaluation survey can be used to follow up more respondents by telephone to resolve problems in the analysis and interpretation of evaluation results. If the search analyst feels that important details on the performance criteria for each database accessed are lost by overall evaluations of citations retrieved from all the databases accessed for searches, users can be contacted by telephone to solicit the data needed for investigating suspected problems or failures in searching specific databases.

Comparisons of Evaluation Results

Although evaluation results are of value in ascertaining instances and causes of search failures, extreme caution must be exercised in comparing evaluation results over time and those achieved by one online service with those of another. Recall and precision measures should not be compared over time or between libraries, because the recall and precision levels attained are greatly influenced by the mix of users' preferences

regarding the level of comprehensiveness desired in search results. An increase in the overall percentage of relevant citations, for instance, might indicate nothing more than a growing preference of users for narrow, rather than broad, searches. Similarly, the comparison is meaningless between the overall percentage of relevant citations achieved by an online service serving a community of undergraduate students with that achieved by one serving technical staff working on the research frontier.

About the only significant comparison that can safely be made is in the number of searches performed over time by an online service or between the number performed by one service and that performed by another; the validity of these simple comparisons, moreover, assumes strict comparability in the definition of the online search.

IMPLEMENTATION OF MODIFICATIONS TO ONLINE SERVICES

Evaluation of online services should not be seen as an intellectual exercise, but rather as a procedure that identifies instances of poor search results and user dissatisfaction and, hopefully, suggests modifications to improve the overall provision of search services. The evaluation process does not end, therefore, with diagnosis of causes of failure.

After determining why poor results are produced by a search, the evidence should be scrutinized to ascertain modifications that might be made in any aspect of the online service. Modifications might be made, for example, in policies, procedures, search strategies, terminal hardware, training of search intermediaries, or holdings of thesauri and other database aids.

Ascertaining modifications that might improve performance

of future searches depends on the ability of the search analyst to determine systematic patterns from analysis and interpretation of evaluation results. If performance data for searches in the same general subject area varies consistently from one database to another, for instance, the search analyst should determine whether the variations stem from variations in indexing practices, the search intermediary's knowledge of the databases, or the availability of thesauri and database aids.

Once a modification is implemented, evaluation forms returned for future searches should be scrutinized to determine whether or not the modification has successfully reduced instances of poor search results. Some causes of poor searches, however, are not under control of the managers of online services: "Such failures, which may be termed data base failures, are due to indexing or vocabulary inadequacies in published bibliographic tools."[23]

Sometimes the analysis and interpretation of evaluation results fail to suggest any modification that can increase performance levels of future searches, but instead suggest general categories of search requests that consistently result in poor search results. This finding is, however, valuable, for it enables managers of online services to emphasize to users posing certain questions the greater likelihood of poor search results; this finding can also be used to discourage users posing certain questions from pursing searches that experience indicates are doomed to failure.

CONCLUDING REMARKS

In providing online services, the goal should be to provide users with information that makes them more productive, creative, or innovative, allows them to make better decisions, and prevents them from wasting efforts in repeating research that is already documented in the literature. Owing to methodological problems in proving that information makes users more productive, facilitates better decisions, or expedites innovation, evaluation of online services is usually focused on the level of user satisfaction. However,

> the ultimate objective of any information storage and retrieval system is not subjective satisfaction as expressed by the user, but improved task performance/problem-solving/decision-making by the user. For too long an uncritical attempt 'to make the user happy' has been prevalent in the library and information services profession.
>
> What is needed instead is an attempt to make the user successful.[24]

Longitudinal Studies of Cost-Benefit Effectiveness

Managers of online services, however, are usually ill-equipped in time and statistical skills to perform the detailed and controlled longitudinal case studies needed to determine the impact of online services on making users more productive. This research needs to be done on a highly selective and well controlled basis by experienced behavioral and information scientists and is germane to evaluating the cost-benefit effectiveness of online services. This research is of less value, however, in generating data of diagnostic

value to the primary goal of evaluating online services: identification of instances and causes of search failures and suggestion of modifications to improve performance levels of future searches.

Evaluation of online services must not be regarded as merely an intellectual exercise or as a goal; instead evaluation must be regarded as a means to the goal of improving the delivery of online services. Evaluation must also be regarded as a process consisting of at least five stages: (1) setting goals for evaluation, (2) designing an evaluation form effective in collecting data useful for the attainment of evaluation goals, (3) conducting the evaluation program with adequate follow-up of non-respondents as well as some respondents, (4) analyzing and interpreting the results of evaluation using information, if necessary, from the entire set of search forms and (5) implementing modifications that analysis of evaluation results suggests for improvement of online services.

Intended Use of Recommended Evaluation Form

The search evaluation form developed by the MARS Evaluation Committee is recommended for use in evaluation by users of search results obtained from retrospective searches of document surrogate databases performed by a search intermediary. Developing an evaluation form to collect data of value in accommodating all possible goals for evaluating online services would have resulted in a form of great length, discouraging users from participating in evaluation. This recommended search evaluation form is appropriate for evaluating online searches provided the major goal of evaluation is to collect data on a

representative group of searches that is of diagnostic value in identifying causes of poor search results and user dissatisfaction. If this goal does not apply, the recommended evaluation form should not be used. If online service managers agree that this goal is of primary importance, but set additional goals for evaluation, then additional questions need to be locally phrased for inclusion on search evaluation forms modified for local use.

Survey of Non-Users of Online Services

A complete evaluation program for online services extends beyond the analysis and interpretation of search results reported on evaluation forms by current users of online services. An evaluation program involving the current users of a service is biased by the exclusion of potential users not currently using the service. Managers of online services need to survey non-users to determine causes for failure to use online services, which could stem from failure to publicize the availability of online services in appropriate media, problems in pricing online services, or in almost any aspect of policies and procedures.

Evaluation of Document Delivery System

The enormous increase in bibliographic access to published works made possible by online services often escalates demand for documents unavailable in the local library collection, sometimes greatly increasing demands for interlibrary loan services. Problems and failures in the entire document delivery system can frustrate the expanded bibliographic access provided by online service; therefore, aspects of the document delivery system also require evaluation.

ACKNOWLEDGEMENTS

As Chair of the MARS Committee on the Evaluation of Services from 1978 to 1980 and as a continuing member of the Committee until the close of the 1982 Annual Conference of ALA, I would like to acknowledge contributions of significance to the Committee's work in designing a search evaluation form for recommendation to the MARS and ALA membership. Donald King provided a good theoretical orientation to the general objectives and problems of evaluating information systems. Mimi Penchansky collected sample evaluation forms in general use in all types of libraries, analyzed them as to content, and listed groups of questions of similar intent to show variations in phrasing; this list formed the basis for Committee deliberations on the selection and phrasing of questions. Signe Larson arranged for a pilot test of an earlier draft of the search evaluation form; feedback from participants in the pilot test improved the final draft of the evaluation form. Sarah Pritchard contributed many useful comments based on her experience with users' evaluation of the SCORPIO system at the Library of Congress. The work of the MARS Measurement Committee, formerly chaired by Carolyn Brown, contributed many useful comments, including the definition of the online search accepted for the evaluation form. Sally Knapp, Danuta Nitecki, Guy Westmoreland, and Pamela Sieving, former and current chairs of the MARS Executive Committee, encouraged and expedited the work of the Evaluation Committee. Finally, F.W. Lancaster's contribution to the literature on the evaluation of information services were of great value during

all stages of the Committee's deliberations.

Individuals serving terms on the MARS Evaluation Committee form 1978 to the 1981 Annual Conference of ALA have included: Judith Bernstein, Richard Blood (Chair, 1978-1980), Doris Bolef, Michael Bruer, Karen Fite, Donald King, Signe Larson, Sandy Menegeaux, Mimi Penchansky, David Pilachowski, Sarah Pritchard (Chair 1980-1982), and Amy Wilson.

REFERENCES

1. "MARS Achieves Section Status!" Messages from MARS No. 3:1 (September 1977).
2. Sally Knapp, "From the Terminal of the Chair," Messages from MARS No. 3:5 (September 1977).
3. "Committee on Evaluation of Services; Preliminary Statement of Objectives," Messages from MARS No. 5:3 (March 1978).
4. "Charge to the MARS Committee on Evaluation of Services."
5. Janice Ladendorf, "Information Service Evaluation: The Gap Between the Ideal and the Possible," Special Libraries 64: 273-74 (July 1973).
6. F. W. Lancaster, "Evaluation and Testing of Information Retrieval Systems," in Encyclopedia of Library and Information Science Vol. 8:240.
7. Ibid., p. 234.
8. Ibid.
9. F. W. Lancaster, The Measurement and Evaluation of Library Services (Washington, D.C., 1977), p.146.
10. Donald T. Hawkins and Carolyn P. Brown, "What is an Online Search?" Online 4:12.
11. Ibid.
12. Katherine Emerson, "National Reporting on Reference Transactions, 1976-78," RQ 16:199 (Spring 1977).

13. Hawkins and Brown, "What is an Online Search?" p.13.

14. Ibid., p. 14.

15. Ibid., p. 16.

16. Lancaster, The Measurement and Evaluation of Library Services, p. 141.

17. Renata Tagliacozzo, "Estimating the Satisfaction of Information Users," Bulletin of the Medical Library Association 65:243 (April 1977).

18. Ibid., p. 246.

19. Ibid., p. 247-48.

20. Ibid., p. 247.

21. Ibid., p. 243.

22. Hawkins and Brown, "What is an Online Search?" p. 17.

23. Lancaster, The Measurement and Evaluation of Library Services, p. 152.

24. Dagobert Soergel, "Is User Satisfaction a Hobgoblin?" Journal of the American Society for Information Science 27:257 (July-August 1976).

APPENDIX: SEARCH EVALUATION QUESTIONNAIRE

Recommended by ALA/RASD/MARS Committee on
Evaluation of Services

SEARCH EVALUATION QUESTIONNAIRE

In order to evaluate and improve computerized literature search services, we would appreciate your taking a few moments to complete and return this questionnaire.

****

Search number_____

Since only a limited sample of users is asked to evaluate the results of their computerized literature searches, the validity of the sample results depends on subsequent follow-up of non-respondents. The search number was entered by a member of the Library staff in order to identify users who return their questionnaires and to eliminate them from a subsequent telephone follow-up of non-respondents. If you prefer to complete and return this questionnaire without indicating your name in order to remain anonymous, the search number will not be used to identify your response. Whether or not you choose to remain anonymous, your response will be kept strictly confidential.

****

NAME:_____

Address:_____

Telephone number(s):_____

Status: (Categories vary by type of library -- local option phrasing)

e.g. Academic Library: Faculty_____ Graduate student_____
Undergraduate_____ Staff_____
Other (specify)_____

e.g. Special Library: Administrator_____ Salesman_____
Laboratory technician_____
Other (specify)_____

1. What was your main purpose in requesting this search? In other words, at the time you submitted your search request, what had you planned to do with the results?

   (Local option -- provide list of possible responses, e.g. term paper, PhD dissertation, faculty research, grant proposal, etc.)

2. Was the purpose of this search to determine that no previous work had been done on this topic?

   Yes_____ No_____

3. Does this search provide enough relevant citations for the purpose for which you submitted the search request?

   Yes_____
   No, but didn't expect to see anything_____
   No (please comment)_____ _____

4. Among the total citations provided by this search, what percentage appears relevant to the specific question or topic for which you submitted a search request?

0%_____
1 to 25%_____
26 to 50%_____
51 to 75%_____
76 to 100%_____

5. Among the total citations provided by this search, what percentage appears relevant to your overall information need, rather than simply relevant to the specific question submitted as a search topic or question?

0%_____
1 to 25%_____
26 to 50%_____
51 to 75%_____
76 to 100%_____

6. Among the relevant citations provided by this search, what percentage is new to you, or, in other words, was unknown to you at the time you examined the search results?

0% of the relevant citations are new to me_____
1 to 25% of the relevant citations are new to me_____
26 to 50% of the relevant citations are new to me_____
51 to 75% of the relevant citations are new to me_____
76 to 100% of the relevant citations are new to me_____

7. Do you feel that the citations that are both relevant and previously unknown to you are worth the cost that you paid for the search?

Yes_____ No_____
If "No," please comment:

(Optional question - may be omitted if library does not charge)

8. Was the time lapse between submitting your search request and receiving your search results reasonable?

Yes_____ No_____
If "No," please comment:

9. Were the results of the search of value to you?

Yes_____ No_____
If "No," please comment:

10. The major reason for the Search Evaluation Questionnaire is to obtain your comments and suggestions for improving the computerized literature searching service. If you have suggestions as to how any aspect of the search service can be improved, please comment in detail: